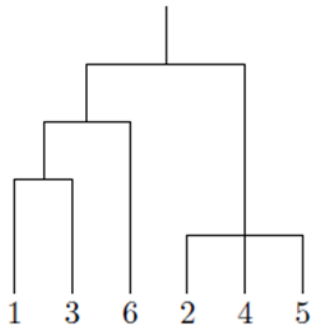
4. Aşamalı (Hiyerarşik) Kümeleme Algoritmaları

Aşamalı Kümeleme Analizi (AKA) algoritmaları da kendi içerisinde Ayrıştırıcı ve Birleştirici olmak üzere ikiye ayrılmaktadır.

Ayrıştırıcı yaklaşımda tüm gözlemler bir kümede olacak şekilde işleme başlanır ve merkeze en uzak/benzemez olan gözlem kümeden ayrılır. Bu şekilde tüm gözlemler tek başına bir küme oluşturana kadar süreç devam eder.

Birleştirici yaklaşımda ise tam tersine öncelikle tüm gözlemler tek bir küme alınarak işleme başlanır ve birbirine en çok benzeyenler bir araya getirilerek yeni kümeler oluşturulur. Bu yaklaşımda da analiz tüm gözlemler tek bir kümede buluşunca sona erer. Özellikle istatistiksel paket programlarda birleştirici yaklaşım kullanıldığından dolayı bu bölümde de sadece birleştirici yaklaşım algoritmaları üzerinde durulacaktır. Ancak algoritmalara geçmeden önce Şekil 4.1'de kümeleme hikayesini gösteren ağaç diyagramı (dendrogram) örneği gösterilmiştir. Birleştirici yaklaşım aşağıdan yukarı kümeleme yaparken, ayrıştırıcı yaklaşım ise yukarıdan aşağıya doğru kümeleme yapmaktadır.



Şekil 4.1 Bir ağaç diyagramı (dendrogram) örneği

Birleştirici aşamalı kümeleme algoritmaları genel olarak şu yapıya sahiptir:

i. n tane birey, n tane küme olmak üzere işleme başlanır ve **D** uzaklık matrisi hesaplanır.

ii. En yakın (d_{AB} değeri en küçük olan) iki küme birleştirilir.

iii. Küme sayısı bir azaltılarak uzaklıklar matrisi güncellenir

iv. (ii) ve (iii) nolu adımlar $n-1$ defa tekrarlanır ve tüm gözlemler tek kümede toplanınca işlemler sonlanır.

Algoritmaların genel işleyişi yukarıda verildiği gibi olmakla birlikte algoritmalar arasında bazı farklılıklar vardır.

Algoritmalar

4.1. Tek Bağlantı (En Yakın Komşuluk) Yöntemi

Tek Bağlantı yönteminde yukarıdaki algoritma uygulanırken uzaklık matrisinin güncellenmesi sırasında herhangi A ve B kümesi arasındaki uzaklık, söz konusu kümelerdeki noktalar arasındaki en küçük uzaklık olarak tanımlanır ve Eşitlik 4.1'de gösterildiği gibi hesaplanır.

$$d_{AB} = \min\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\} \tag{4.1}$$

Burada $\boldsymbol{x}_i$ ve $\boldsymbol{x}_j$ sırasıyla A ve B kümesindeki elemanları ifade etmekte olup, $i = 1,2,\dots,n_A$ ve $j = 1,2,\dots,n_B$'dir.

4.2. Tam Bağlantı (En uzak komşuluk) Yöntemi

Bu yöntemi tek bağlantıdan ayıran, uzaklık matrisinin güncellenmesi sırasında en küçük değil en büyük uzaklığı almasıdır. Yani algoritmanın (iii) adımında uzaklık matrisi güncellenirken A ve B kümeleri arasındaki uzaklık Eşitlik 4.2'deki gibi güncellenir.

$$d_{AB} = \max\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\} \tag{4.2}$$

4.3. Ortalama Bağlantı Yöntemi

Bu yöntemde de uzaklık matrisi güncellenirken A ve B kümeleri arasındaki uzaklık A kümesinde bulunan n_A ve B kümesinde bulunan n_B gözlem arasındaki uzaklıkların ortalaması olarak alınır ve Eşitlik 4.3'te gösterilmiştir .

$$d_{AB} = \frac{1}{n_A n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{4.3}$$

4.4. Küresel Ortalama (Centroid) Yöntemi

Bu yaklaşımda A ve B kümeleri arasındaki uzaklık güncellenirken centroid olarak adlandırılan iki küme merkezi (kümelerde bulunan gözlemlerin ortalama vektörleri) arasındaki Öklid uzaklıkları kullanılır. Bu tanımdan hareketle söz konusu uzaklık Eşitlik 4.4'te tanımlanmıştır.

$$d_{AB} = d(\overline{\boldsymbol{x}}_A, \overline{\boldsymbol{x}}_B) \tag{4.4}$$

Eğer uzaklık matrisindeki en küçük uzaklık A ve B kümeleri arasında ise bu kümeler birleşecektir. Bu yaklaşımda kümenin merkezinin hesaplanması gerekir. Bu birleştirilmiş kümenin merkezi ağırlıklı ortalama olarak Eşitlik 4.5 verildiği gibi hesaplanır.

$$\overline{\boldsymbol{x}}_{AB} = \frac{n_A\overline{\boldsymbol{x}}_A + n_B\overline{\boldsymbol{x}}_B}{n_A + n_B} \tag{4.5}$$

4.5. Ortanca (Medyan) Bağlantı Yöntemi

Eğer A ve B kümeleri centroid yöntemi ile bir araya getirilirse ve A kümesinin eleman sayısı B kümesinden fazla ise Eşitlik 4.5 ile hesaplanan $\overline{\boldsymbol{x}}_{AB}$ merkezi, $\overline{\boldsymbol{x}}_A$'ya $\overline{\boldsymbol{x}}_B$'den daha yakın olacaktır. Dolayısıyla küme sayısının ağırlığını önlemek amacıyla diğer kümeler ile birleştirilmiş AB kümesinin uzaklığını hesaplarken kullanılacak yeni merkez, ortalama yerine ortancaya/orta noktaya (median/midpoint) dayandırılır. Buradaki ortanca ifadesi istatistikte sıklıkla kullanılan medyanla birebir aynı değildir ve şöyle tanımlanır:

$$m_{AB} = \frac{1}{2}(\overline{\boldsymbol{x}}_A + \overline{\boldsymbol{x}}_B) \tag{4.6}$$

Özetlemek gerekirse Centroid ve ortanca yöntemleri arasındaki tek fark, birleştirilmiş iki kümenin yeni merkezinin hesaplanmasındaki görüş farklılığıdır.

4.6. Ward Yöntemi

Ward yöntemi küme içi karesi alınmış uzaklıkları en küçük, kümeler arası karesi alınmış uzaklıkları da en büyük yapmayı amaçlar. Öncelikle A ve B kümeleri için kare toplamları;

$$KT_A = \sum_{i=1}^{n_A} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_A)'(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_A) \tag{4.7}$$

$$KT_B = \sum_{j=1}^{n_B} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_B)'(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_B) \tag{4.8}$$

şeklinde tanımlanır. Eğer A ve B kümeleri birleştirilirse elde edilecek yeni kümenin kareler toplamı da Eşitlik 4.9 ile hesaplanır.

$$KT_{AB} = \sum_{k=1}^{n_{AB}} (\boldsymbol{x}_k - \overline{\boldsymbol{x}}_{AB})'(\boldsymbol{x}_k - \overline{\boldsymbol{x}}_{AB}) \tag{4.9}$$

Burada $\overline{\boldsymbol{x}}_{AB}$ Eşitlik 4.5'te verildiği gibi hesaplanır. Ward yönteminde A ve B kümesi birleştirildiğinde kareler toplamındaki artışın mümkün olduğunca düşük olması istenir. Bunu ölçmek için Eşitlik 4.10'da gösterilen değerin en küçük olması istenir. Tüm mümkün durumlarda en küçük I_{AB} değerine sahip kümeler bir araya getirilir.

$$I_{AB} = KT_{AB} - (KT_A + KT_B) \tag{4.10}$$

KAYNAK

Hasan Bulut (2018). "R Uygulamaları ile Çok Değişkenli İstatistiksel Yöntemler", Nobel Akademik Yayıncılık, Ankara.